A STUDY OF THE EXAMINER RELIABILITY IN ESSAY- TYPE EXAMINATION (SUBJECT HISTORY)

by

ARUN KUMAR GHOSH

DEPARTMENT OF PSYCHOLOGICAL FOUNDATIONS

(NATIONAL INSTITUTE OF EDUCATION)

NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

1964 - 65

ACKNOWLEDGEMENT

I have great pleasure to place on record my heartfelt gratitude and sincere thanks to Dr. Shib K. Mitra, Head of the Department of Psychological Foundations (National Institute of Education) for the encouraging help given to me by him.

I acknolwedge my indebtedness to Dr. S.S. Kulkarni, Dr. Harold Webstar, Member of the Teachers College Columbia University team from U.S.A., Dr. A.B.L. Srivastava and Mrs.Shukla for their help in designing the experiment and interpretation of the data.

Mr. Om Prakash, Lecturer in C.I.E. and Mr. Kanti Chandra Dutt of Psychological Foundations have structured the questions. I owe my thanks to them.

My heartiest thanks go to Dr. Prayag Mehta, Mr. C.M.N. Mehrotra, my guides, for their constant encouraging help and to Mr. R.K. Mathur, my guide and convenor, whose immaculate friendship and assistance at every step starting from designing to interpretation provided inspiration throughout the progress of the report.

I shall be failing in my duty if I fail to acknowledge my thanks to the Principals and teachers who extended their unstinted co-operation in getting the script valued in time unmindful of the inconvenience caused to them. My thanks and appreciation are also due to the boys who took the test.

The constructive criticism of my friends in the research Methodology course meant much to me at the preparatory stage. I owe my thanks to them.

ARUN KUMAR GHOSH

TABLE OF CONTENTS

1.	INTRODUCTION	1
2.	REVIEW OF THE PAST STUDIES AND THE PROBLEM	4
3. a.	PROCEDURE	13
b.	DESIGN OF THE EXPERIMENT	17
4.	STATISTICAL ANALYSIS OF DATA	20
5.	DISCUSSION AND INTERPREATION	24
6.	CONCLUSIONS	29
7.	SUMMARY	31
8.	REFERENCES	33
9.	APPENDICES	35

INTRODUCTION

The purpose of Examination is to assess the extent of achievement, an outcome of the process of education. " No field of human endeavour, whether cultivated by public or private agencies, is more exposed to criticism than that of education " and "within the education field the sector which receives and probably deserves most criticism is the examination system"[1].

During the last few decades, the examination system has come under heavy criticism. All concerned with education have questioned the utility and reliability of the present system of examination. Various Education Commissions and Committees appointed by the Government of India at different times have made many recommendations for improving the Examination System both at the school and college levels. We can guess the alarming condition of the present system from the following remark of the University Education Commission of 1948 - if we are to suggest one single reform in university education it should be that of examination. The Secondary Education Commission of 1952 also suggested a thorough change of the present system of examination both internal and external.

The immediate effect of all these criticism was that several research projects on the reform of Examinations were taken up by the Ministry of Education and various other educational departments. Dr. Edwin Harper (Jr), Dr. H.J. Taylor, Dr. A.K. Gayen and others are working for the improvement of

---

(1) Preface by John Surgeant to "Measurement of Achievement English", Report II by Dr. A.K. Gyar, R.K.Mathur and others.

the examination system.

"Examination is a necessary evil". It has come to stay. 'Abolish all examination' - a slogan of the extremist reformer can not be entertained. We must mobilise all our efforts to find a way out from the present circumstances. But how to do it? We can probably achieve the goal of examination reform by improving its reliability and validity. "Examination include oral examination, conventional written examinations or tests, objective examination and standardised tests" (14) and essentially concerns pupil & teachers in the capacity of a teacher, a paper-setter and an examiner.

The present study was concerned with the examiner particularly Examiner - reliability in essay-type examination.

The concept of reliability is involved in every kind of measurement method. Reliability necessarily means the accuracy with which a measuring device measures whatever it is expected to measure. "An ideal measuring instrument would yield the same results everytime it was used and the same results regardless of the particular form of the test involved" (21). Of course perfection can not be obtained in essay type examination. So we should seek some method where the scores are least contaminated with errors.

Reader reliability or examiner reliability may be operationally defined as the correlation between the series of marks awarded by two or more examiners independently to the same answers. It is with this point in view the present study was designed.

I am aware of the fact that any reform of examination system is hardly possible without taking into account the entire education process starting from the formulation of objectives of education at the begining to the evaluation of the effects at the end, but for the purpose of the present study I have not tried to manipulate the present educational process and tried to study the examiner -reliability under specific directions and no direction with regard to i) questions as is usually set in the higher secondary examination and ii) structural questions, a modified form of the usual questions.

REVIEW OF RELATED STUDIES AND THE PROBLEM

Before giving a review of the related studies, it will be worthwhile to have a few lines regarding reliability itself. Reliability refers to consistency in measurement. Regarding test reliability there are four major methods of estimating test-reliability. These are i) parallel or equivalent form or alternate form method, ii) the test -retest method, iii) the split-half method and iv) the analysis of variance method. But these methods can not be applied to present system of examination as there is only one examination. Neither there is any parallel form nor it is duplicated.

The estimation of reliability of essay type examination creates an unique problem. Cureton (15) has said "that for an objective test there are just two measurable types of unreliability: that associated with sampling in the universe of possible items, and that associated with sampling in the universe of possible reactions of the same individuals to the same items on different occasions. For a subjective test ....there are two more : that associated with sampling in the population of possible judges, and that associated with sampling in the universe of evaluations of the same papers by the same judge on different occasions." It is the last two that we are concerned with. Cureton has rightly pointed out that in objective type of test we are not bothered about the consistancy of scoring. It can also be done by a machine.

The question of subjectivity in marking is very much involved in essay-type examination. In subjective examinations the concept of reliability breaks up into two distinct parts i) reliability of marking or examiner-reliability and ii) reliability of the essay itself.

Examiner-reliability may be taken as the correlation among marks.given by different examiners. The accuracy with which the same examiner or different examiners will grade the answer-script is a major problem in essay-type examinations. There are a number of studies on examiner-reliability. Inconsistency of marking was the overall findings of these studies. These studies were mainly concerned with various system of scoring or marking or evaluation.

Among these studies that done by Edgeworth is the earliest. He found that the marks given by twenty-eight examiners on a piece of Latin prose varied from 45 to 100, the total mark being 100. He listed several causes like the examiner's idiosyncracies, negligiance of examiners regarding scale etc. for this variation in marking.

Starch and Elliott in 1912-13 conducted a study in America. In their study they got two answerscript of English, written by high school students, examined by 142 examiners. They found variation as large as 35 to 40 points. A similar investigation was carried out on Geometry and History marked independently by 70 examiners. On Geometry, the variation was found 28% to 92%. History also showed similar results. The investigations conducted by International Institute Examinations Enquiry Committee, Hartog, Rhods pointed

clearly to marked discrepancy among different examiners. Subjects like mathematics & chemistry were no exception. Vernon & Millican (16) conducted a study on reliability on English essay and found examiner reliability of 51. They concluded on the basis of their study that 4 factors probably account for most the variation 1. The less skilled or mature the writers or students, the more their essays are marked for mechanics rather than for the relatively subjective qualities of style and thought content -2. Some markers are more experienced than others -3. A long essay may be more objectively assessed than a short one.4. Fourth factor is the heterogenity or range of ability among the writers.

A.D. Datta made a detailed investigation of essays written in Bengali. 30 essays were marked by 24 teachers. He studied both variation in marking and also inter-reader agreement. The most surprising, finding was that a paper ranked from 1 to 30 and it had the highest average mark. Fifteen of these answer-scripts were re-examined after five year. The result was equally disappointing. In another study Dutta (18) investigated the variation of marks given by four independent examiners in English, history, geography and mathematics by four examiners independently. The inconsistency of marking was predominant here also. The examiner-reliability was highest in mathematics and lowest in English and history.

Thomson and Bailes conducted a study in which they found the correlations between marks of six experienced examiner vary from 42 to 85. P.H. Taylor (8) studied the effect of

three different instructions (Analytic, Devon, personal) to markers or examiners. He found that instruction to markers have no significant effect on the marking. The difference which do exists are due either to chance errors or to interaction with other variables. The result suggested that against a marking background of general impression, radical differences in instruction to markers make no general difference to the marking or to the reliability of examiners. Hartog (6) conducted a study on general impression vs analytic marking. The analytic method was in terms of quantity, quality and control of ideas, spelling, grammer punctuation, structure of sentences, vocabulary and hand-writing. Ten examiners examined 75 papers by the analytic method and another set of 75 papers was examined by ten examiners by the impression method. The findings did not conclusively proved the superiority of the analytic method over the impressionistic method, though the variability among examiners was somewhat lowered by the analytic method.

Edwards Penfold (11) compared the consitency of marking of both long & short essays. Sixteen examiners marked 165 answerscripts on brief compositions. It was seen that there exists significant variance between examiner's standard and also between the marks of the same examiner on two occasions.

In the second experiment 45 answer scripts of long essays were marked according to a very detailed analytic scheme. Here also the examiner difference was found to be significant. Variation due to different essay topic was also noticed in this study.

A. Yates & D.A. Pidgeon in a study used a complex design. 245 answer-scripts was marked by seven examiners. It was seen that the examiners differ significantly in mean & standard deviation.

H. Lamb (20) in his study (Thesis) compared three marking schemes. Eight teachers marked 40 composition script from a selection examination. The scripts were remarked again after four months. The three methods of marking were:

i) impression marks given by four teachers, and analytic marks given by each of eight markers. It was found that for the impressionistic method marker agreement was ·736 and mark-remark correlation was ·861. For the analytic method marker agreement was ·861 nad mark-remark reliability was ·907. In teamed impressionistic method examiner reliability was ·918 and mark-remark reliability was .961. The above result shows that the teamed impressistic marking is better than the other two and analytic method is better than individual impressistic method.

There are several studies on Essay reliability by Penfold, Vernon & Millican, Wiseman and Wringley. Comparison between reliabilities for English essay and objective achievement tests were also made. Studies by S. Wiseman, A.E.G.Pilliner, A. Yates and D.A. Pidgeon and D.S. Finlayson showed that the reliability for English essays range from .72 to ·89 and reliability for the objective test range from ·88 to ·95.

Stalnaker referred to a case where an essay question written by 6384 students were marked independently by a number of experienced teachers twice. Each paper was classified according to examiner's judgement. The total number of

categories was eight. It was found that in only 41 % of cases the grade given by the second examiner tallied with the grade given by the first. Sims in a study showed that there is possibility to improve reliability of Essay examination through "graded by rating". He showed that the objectivity co-efficient was higher than that obtained through grading by percentage.

In a study by Eells answer-scripts on school history and geography were examined by sixty-one teachers twice. The time gap between the two evaluation was eleven weeks. For a particular question only ten out of sixty-one teachers gave the same grades on two occasions.

There was one important study by cast (2) where 40 essays written by central school girls of 14 were judged according to four methods ; i) Depending on examiner's discretion, ii) the achievement of aim method by Hartog, iii) the general impression method and Burt's analytic method correlational and variance analysis showed a great variability among the examiners. Though analytic method was the best, the general impression method was also very near to it and nothing could be definitely said about the superiority of the analytic method.

A.K. Gayen in India made a rigorous statistical analysis of answer scripts in school final, Mathematics & English of West-Bengal. The split half reliability co-efficient was far below than that of an objective test. The Directorate of Extension Programmes for Secondary Education is also working on improvement of the present system examination. They have

published a few booklets like

i) Improving examinations (By P.D. Sharma) in which they have dealt with need, considerations & characterstic of a good examination and forms of questions.

ii) The concept of evaluation in education dealing with objectives, learning experiences, evaluation devies & procedures etc.

iii) Evaluation in General Science dealing with the objectives of teaching general science, developing test material etc.

iv) Evaluation in Mathematics dealing with the need for a change in the pattern of examination in Mathematics, objectives of teaching mathematics, developing test materials.

Dr. H.J. Taylor conducted study on two or three examiners in 1962. The examiners agreed in 4.8 %. cases only in assigning total marks. The Central Examination Unit of DEPSE (NCERT) has published a monograph giving the list of specific direction given to the examiner's by the different examination Board. However, no rigorous study seems to have been undertaken in India to assess the effectiveness of these directions in stepping up examiner-reliability. The present study is a modest effort in this regard.

The Problem:

The greatest draw back in the present-day examinations is its low reliability. We set apart for the present the criticism that question papers in public examinations generally measure cramming and not that of real understanding. What is more important is the difference in the frame

of reference among examiners. It has been seen that certain examiners are in the habit of giving high marks where certain examiners give consistently low marks and this discrepancy among examiners is very serious as it directly effects the career of the students through misclassification. The lack of inconsistency of marking may be due to several factors. The language of the question may be such that different examiners may take different views about the right type of answer. The examiners generally give a round or absolute mark on the answer of a question, though the question may have two or three aspects to write. As the round mark on the question is not split into as many parts as there are aspects, examiners assessment can hardly be uniform. Because of this lack of uniformity in marking a student who truly belongs to a given ability group stand the chance of being misplaced on the basis of unreliable scores in some other ability group. Due to uncertainity in marking, a student whose true score is 45 may get 38 or 39 & loose his second class. A most pitiable case is that where a student has got good marks in all the papers but failed in one paper by 1 mark, he would have well passed in this subject had it been examined by another examiner, but for this one mark he had to waste one year. Similarly is the case of those who miss good class by 1 or 2 marks. As in India the class or division obtained in public examination is given undue weight, the unreliable scores often plays a devasting role & spoil the whole career of a student.

In our effort to improve the scoring system we are most concerned with the actual mark given and not much with the rank. I like to explain the idea why I do not like to give

more weightage to rank. In public examination ranking is strictly done or is most useful only 100 or 50 cases i.e. those who are eligible or contestant for scholarships. But overall marks effects each and every individual. Let us take an Example: A & B are two examiner examining three scripts. The whole thing is represented in tabular form.

	A	B	
1	65	57	Here the rank given
2	55	44	by the two examiners
3	36	28	are the same, but

in one case A got first division, the other placed him in second division. In the third case, one examiner has got the candidate failed.

Two examiners may give a student the same rank, but the absolute marks given by them matter much in most cases and fetches good class. Hence our problem is how to reduce the difference in means of the marks given by examiners. A boy who gets 60 from one examiner, should get near about the same mark from any other independent examiner on the same script and on different occasions. This is the problem.

PROCEDURE

Reliability is a widely used concept and reliability of essay type examination pose two unique problems as mentioned by Cureton (15). As questions of History set at the Higher Secondary final examination are of subjective-type, there are two specific causes of unreliability. One is associated with the evaluation of the same paper by the same judge or different judges on different occasions. The other is associated with the items. The present study was confined to the former one.

It happens generally that the answerscripts of Higher Secondary examination are collected and sent to different examiners. These examiners may vary in academic qualifications and also in teaching experiences. They may be either from among the school teachers or college lecturers. Each answer-script is examined by only one examiner. No device is employed to compare the marks given by examiner with another examiner as different examiners examine different answer-script. Of course, there is a Head-examiner who sometime instructs other examiners regarding the method of scoring. Head-examiner may also scrutinise a few papers marked by each examiner. But all these measures do not help much. Different studies has shown the alarming discrepancy between the marks given by different examiners. One particular examiner may consistently give high marks and another may consistently give low marks and these in turn effect the over all total as well as grade of the pupil. Not only this one pupil who

failed, may have passed, if the script had been examined by an examiner who is in the habit of alloting high marks.

The present study was designed to find a method where the discrepancy among the marks given by different examiners will be less. The subject chosen was History as in this subject there are lesser number of studies and at the same time this subject suffers from subjectivity most.

For this study questions were selected from the questions already set in higher secondary final examinations. in Delhi. The question were from the set of questions set in 1963 & 1964. There were three questions in all. Two questions were from the question paper of 1963 and the third one was from that of 1964. There were two types of questions. One was exactly identical with that set in the usual examinations. The same questions were structured or broken under several heading/ by a Research Associate of Psychological Foundations and a Lecturer in History of the Central Institute of Education. The questions were the following:

1. What influence did Muslim contact/have on Hinduism?
2. Discuss the causes and nature of risings of 1857?
3. Evaluate critically the chief reforms of Cornwallis.

In another type these questions were given to students in the following form:

1. What influence did Muslim contact have on Hinduism?
   Discuss under the following headings:
   a) Art, Mu sic & Painting
   b) Architecture (Buildings)
   c) Dress and customs (culture)

d) Literature

e) Religion (Workship)

2. Discuss the causes and nature of the risings of 1857 under the following heading:

a) Social

b) Religions

c) Financial

d) Military

e) Military revolt or Ist war of independence

3. Evaluate critically the chief reforms of Cornwallis under the following headings:

a) Reform of public services

b) Judicial reforms

c) Police reforms

d) Permanent Settlement of Bengal (Land revenue reform).

Each question carried 15 marks. Students were asked to answer all the questions. Total time allotted was one and a half hour. Sample of students were from the population of students appearing at the Higher Secondary Examination 1965.

Examiner's were all Senior teachers of History of Higher Secondary Schools in Delhi. One set of examiners were given a few instructions, others were not given any instructions. The instructions given to examiners were in the form given below. Please try to follow the following directions:-

1. Please do not put any mark whatsoever on the answersheets.

2. Think of a model answer before examining the answerscript.
3. Read all answers to one question before going on to the next.
4. Do not give stress on language. Do not deduct marks for spelling mistakes, punctuation etc.
5. Avoid giving negative marking.
6. Please keep in mind the different weightage given to different parts of the questions as given below.

Q.1. What influence did Muslim contact had on Hinduism.

Art, music & painting	3 marks
Architecture (Buildings)	3 marks
Dress & customs	3 marks
Literature	3 marks
Religion	3 marks

Q.2. Discuss the cause & nature of the rising of 1857.

Social	3 marks
Religions	3 marks
Financial	2 marks
Military	2 marks
Military revolt or Ist war of independence	5 marks

Q.3. Evaluate critically the chief reforms of Cornwallis.

Reform of Public Services	3 marks
Judicial reforems	3 marks

Police reforms	2 marks
Permanent Settlement of Bengal	7 marks

7. Each question carries 15 marks.

Instruction No. 5 needs a bit of explanation. By 'negative marking' the experimenter meant that sometime it so happens that the examinee writes so absurd and nonsensical things that the examiner gets irritated and the result may be that the examiner may deduct a few marks which he would not have, had the examinee not written those absurd things. The term 'negative marking' was explained to the examiners when the experimenter met them before handing over the answer-script. All the answerscripts were written in Hindi.

DESIGN OF THE EXPERIMENT

The design of the experiment was of 2 x 2 factorial design type. Altogether twenty boys were selected from three different schools. They were divided into two groups by matching them against their achievement in History at the last annual school examination i.e. the criterion of matching was their ws mark in History at the annual examination of class X. One group of boys was given questions as usually set in the final examination. The other group was given the structured questions. Each group of boys was again broken into two groups with the help of random number table. Diagramatrically the design can be

represented as

	Examiner	
	No Direction	Direction
Unstructured	Gr 1	Gr 2
Structured	Gr 3	Gr 4

In all there were four groups of students and each group contained 5 students each. Each answerscript was examined by 5 (five) examiners. So in all twently examiners were selected randomly. They were all post graduates in History, teaching in Higher Secondary Schools in higher classes. The selection of examiners was done in the following manner:

1. From the list of schools supplied by the directorate of Education, Delhi. Nearly about 44 schools were selected by the help of random number table.
2. The Head of the Department issued letters to the principals of these schools requesting them to extend the assistance of their senior History teachers. A stamped self-addressed card was attached to every letter for their early reply. Twenty two schools returned the card in time expressing their desire to help.
3. The allocation of teachers to different groups was again done with the help of random No. table.

Each examiner had to go through 5 answer-scripts i.e. all the answerscripts of one group. The sample size had to be kept small because of the paucity of time and lack

of fund. The work done by the examiners was purely voluntary.

The data was written in tabular forms[1]. Marks given on each question was also tabulated. The analysis of data was was done by applying i) Analysis of variance with raw as well as adjusted data ii) Intra-class correlation and iii) co-efficient of concordence. For the total marks on each script all the three tests were applied.

---

1. See Appendix.

ANALYSIS OF THE DATA

The marks given by examiners were written in a tabular form shown below:

Table 1 Group 1.

	E_1	E_2	E_3	E_4	E_5	Total
A	27	25	19	24	23	118
B	18	24	15	17	14	88
C	19.0	24	17	16	14	90
D	15	20	14	12	7	68
E	21.0	21	10	19	10	81
Total	100	114	75	88	68	445

A, B, C, D, E are the five different answerscripts and E_1, E_2 .. E_5 are the five different examiners. The column represents the score given by the examiners on different scripts and the rows represent the scores given by five examiners on the same script.

The Analysis of variance table for Group 1:[1]

Table 1.1

Sources of variation	df	ss	ms	
Between student	4	269.6	67.4	$F = \frac{69.2}{6.1}$
Within student	20	374.4	18.72	$= 11.34$
Between Examiner	4	276.8	69.2	
Residual	16	97.6	6.1	
Total	24	644		

F - ratio is significant at .01 % level. This shows that the examiners differ significantly from each other.

Intraclass correlation was found out by applying the formula

$$r_e = \frac{\text{ms between student} - \text{ms within student}}{\text{ms between student} + (K-1)\ \text{ms within student}}$$

In this case $^1r_c = \frac{67.4-18.72}{67.4+(5-1)\ 18.72}$

$$= \frac{48.68}{142.28} = .35$$

Difference in scoring may be due to two factors 1) Error of measurement 2) difference in frame of reference. As the intraclass correlation was found to be low, the great differences in the frame of reference of the examiners was adjusted to see if the intraclass correlation goes high. Again we take the case of Group 1.

	E_1	E_2	E_3	E_4	E_5	Grand Total
Total marks	100	114	75	88	68	445
						G=17.8
Mean	20	22.8	15	17.6	13.6	
Deviation from G	2.2	5	-2.8	-.2	-4.2	

Table 1.2

Adjusted Data.

	E_1	E_2	E_3	E_4	E_5	Total
1.	24.8	20	21.8	24.2	27.2	118.0
2.	15.8	19	17.8	17.2	18.2	88.0
3.	16.8	19	19.8	16.2	18.2	90.0
4.	12.8	15	16.8	12.2	11.2	68.0
5.	18.8	16	12.8	19.2	14.2	81.0
Total	89.0	89.0	89.0	89.0	89.0	445.

1. r_c^1 stands for Intraclass correlation.

Table 1.3

Analysis of Variance Table with Adjusted Data.

Source of variation	df	ss	ms
Between student	4	269.6	67.4
Within student	16	97.4	6.09
Total	20	367.0	

r_c = Intraclass correlation = $\frac{67.4 - 6.09}{67.4 + 4(6.09)}$ = .68

So the intraclass correlation went higher by making adjustment for difference in frame of reference of different examiners. Adjustment for the frame of reference may be defined as the deviation of the mean rating of an examiner from the mean rating of all examiners.

The co-efficient of concordence was also found out for all the 4 group to see the agreement between examiners regarding ranks given to each student. The formula used was (due to Kendall) $W = \frac{12\ S}{m^2\ N\ (N^2-1)}$ Where W is the co-efficient of concordence.

S = The sum of squares of the deviations of m sums -of-rank around their mean.

m = number of Examiners or judges.

N = Number of subjects.

Calculations for Group 1:

Table 1.4

	E_1	E_2	E_3	E_4	E_5	Sum of ranks	Square of sum of ranks
1	1	1	1	1	1	5.0	25
2	4	2.5	3	3	2.5	15.0	225
3	3	2.5	2	4	2.5	14.0	196
4	5	5	4	5	5	24.0	576
5	2	4	5	2	4	17.0	289
						75.0	1311

$$S = 1311 - \frac{75^2}{5} = 1311 - 1125 = 186$$

$$W = \frac{12 \times 186}{5^2 \times 5 \times (5^2-1)} = \frac{186}{50 \times 5} = .744$$

The maximum difference among the examiners was also found out for each group and the average of maximum difference on the five scripts was calculated. The whole thing can be summarised as below:

Table 5.

Group	F-ratio	Interclass correlation raw scores	Adjusted	Co-efficient of concordence.	Average range.
1	11.34	.35	.68	.744	10.8
2	6.276	.47	.669	.728	4.8
3	16.8	-.03	.344	.212	16.0
4	3.924	.17	.46	.406	7.5

DISCUSSION AND ENTERPRETATION

The F - ratio[1] shows that the disagreement among examiners is very large. Only in the 4th group the F-ratio is not significant at 1% level.

The variability in evaluation of the same script by independent evaluators may be because of i) difference in the frame of reference of the examiners. This is not of a random nature, and can be ironed out by suitable scaling devices like the method of equivalent scores or t -scores etc; ii) error of measurement, this is of a random nature. If there is little error of measurement and the differences in mark is mainly because of difference in examiner's frame of reference (implicit norm adopted by the examiner) the co-efficient of concordance or intraclass correlation may show higher value while the F-ratio will be significant or the difference between maximum & minimum mark may be still high. If the intraclass correlation is low and the F-ratio is insignificant, the interpretation may be that the inconsistency in independent evaluation of the same script is mainly because of error of measurement and not as much due to difference in frame of reference.

In the case of Gr 1 we find that F-ratio is highly significant. It shows that the examiners differ significantly in scoring but the ranks assigned to student by different

---

1. See table 5 in Procedure.

examiners do not differ significantly as we get a high value of co-efficient of concordance (.744). The intra-class correlation .35 is also higher in this group than that of Gr 3 & Gr 4. Group 2 also shows similar result. Here F-ratio is significant but co-efficient of concordance is high (.728) and reliability is a little higher than Group 1 and much higher than Gr 3 & Gr 4. The average of the differences in maximum and minimum marks is the lowest (4.8). If we compare Gr 1 and Gr 2 we find that the only effect of direction has been to reduce the range between maximum and minimum marks. In both Groups 1 & Group 2, the intraclass correlation & co-efficient of concordance are similar but the F-ratio is significant, though in one case it is 11.34 and in Gr 2 it is 6.276. If we partial out the effect of examiner's difference due to frame of reference, the intraclass correlation becomes almost equal (.68 & .669).

In Group 3, where structured question were given to the students and no directions were given to examiners, the result is quite alarming. The F-ratio is highly significant (16.8), the intraclass correlation is very low and co-efficient of concordance is only .212. In this case neither the examiners agree in the mean scores nor in regard to ranks given to each student. The average range between maximum & minimum marks is also alarmingly high. It is 16.8 and in one particular case the difference is 19 out of a total mark of only 45.

In Group 4, where stmuctured questions were given to the students and directions were given to the examiners,

hopeful results were obtained. Here F-ratio (3.924) is insignificant at α=.01 level. The examiners here do not differs significantly in the mean scores given to students. But the co-efficient of concordance and intraclass correlation is low and this shows that the examiners do not agree in ranks.given to the student. An insignificant F-ratio & low intraclass correlation tells that error of measurement is much but error due to differences in frame of reference is less. The average range between the maximum and minimum mark in this group is 7.5, much lower than that of Gr.3. Here also it points out that the directions help in bringing down the range between the maximum & minimum marks given on the same script.

After partialling out the effect of examiner's differences due to frame of reference the intraclass correlation for Gr 3 goes up from -.03 to .344 and in group 4 it goes up from .17 to .46, far below than that of Gr 1 and Gr 2.

If we compare the result of Gr 1 with Gr 2 (in both the group unstructured questions were given to the student) and Gr 3 with Gr 4 (in both the group structured questions were given), we find that the effect of direction in the presence of structured questions is more. The average range drops from 16.0 to 7.5 (a drop of 8.5) where as in the other case it is 10.8 to 4.8 = 6 only.

In the 4th group both structured questions and directions jointly helped to bring down the difference in mean score given by different examiners. The result somewhat

supports the finding of Hartog (16) who in a study on impression vs analytic method found that the variability among examiners was somewhat lowered by the analytic method.

P.H. Taylor (8) found that instruction to examiners have no significant effect on the marking. The result suggested by him was that against a background of general impression, radical differences in instruction to markers make no general difference to the marking or to the reliability of examiners. The present study also points out to the same opinion. In none of the groups, directions could make any difference. Direction seems to have no effect at all except that it help to minimise the difference between the maximum & minimum mark given on a script.

Impressionistic method is as good as analytic method, the result shows this in the case of Gr 1 and Gr 2. Cast (2) also found similar result in his study. Impressionistic method is better so far the ranking of the student is concerned. Different examiner tend to assign the same rank to each student In analytic method (method with directions) different examiners tend to give differnt ranks to each student. One plausible rival hypothesis to explain this difference may be that due to sampling error (particularly because the sample is small) the students in Gr 1 may be more differing in their knowledge and discrimination among students become easy whereas in the Gr 4 the students are of the same calibre and the discrimination became difficult.

If we do not consider exclusively ranks i.e. relative position of a student among others (we have good reason on to neglect rank which plays an important part only in the case 1st 100 or 50 boys)[1], structured questions with directions give the best result.

---

1. See The Problem on page

CONCLUSIONS

The purpose of the study was to see the effect of structuring i.e. breaking up the main question under several small headings and the effect of directions on consistency of marking.

The study showed that against a marking background of general impression, instruction to Examiners make no general difference to the marking or to examiner-reliability. Directions to examiners helped only to brining down the difference between the maximum and minimum mark awarded on a script.

Another finding was that impressionistic method is better than the method with directions, so far ranking is concerned. Of course, nothing can be said definitely in cases where finer discrimination is needed.

If the absolute mark is taken into consideration, the method which provides directions to the examiners and structured questions to the students is superior to other treatments. The agreement between examiners on scores, not necessarily rank is very important, it is the absolute scores which is very much responsible for misclassification. Hence in the light of this study it could be said that structured questions with directions will be a better substitute for the a usual type of questions with method of scoring by the impressionistic method.

The study suffers from external validity because of its small sample size and motivational factor on the part of the teachers whose work was purely voluntary. The sample of teachers may be treated as different as they are more co-operative. The reason for impressionistic method to be better may be because of that here the examiners are able to take a wholestic view of the answers.

It will be worthwhile to conduct a study with two or more different kinds of structured questions and two or more different type of directions and see the effectiveness of each treatment combinations.

SUMMARY

The study was an attempt to seek an answer to the question whether questions broken under small heads along with directions given to examiner can help to improve examiner-reliability.

The design employed was of 2x2 factorial type. Diagramatically it stands as follows:

	No Directions to Examiners	Directions to Examiners
Usual questions	Gr I	Gr II
Structured questions	Gr III	Gr IV

Four groups of students were taken by matching them against their achievement scores in History at the last annual examination i.e. of class X. Twenty senior history teachers were selected. Allocation of teachers to different groups was done by random number table. Analysis of the data was done by i) analysis of variance both on raw score and adjusted data (after eliminating the difference due to examiner's frame of reference), ii) by computing intraclass correlation and iii) Kendall's co-efficient of concordance.

F-ratio was found to be insignificant at 1% level in the case of Gr 4, but the co-efficient of concordance and intraclass correlation was high in the case of Gr 1 and Gr 2. The results may be put in a tabular form.

Group	F-ratio	Intraclass correlation raw score	adjusted	Co-effic-ient concor-dance.	Average but maxi-mum & minimum mark.
1	11.34	.35	.68	.744	10.8
2	6.276	.47	.669	.728	4.8
3	16.8	-.03	.344	.212	16
4	3.924	.17	.46	.406	7.5

Except Gr 4, in all the other group examiners differed significantly from each other regarding absolute score of the students. Result showed that only direction to examiner do not make any difference in improving reliability of scoring except that it tends to minimise the difference between maximum and minimum marks given ona script. Structured questions along with directions to the examiner was found to be good substitute for the usual type of questions along with impressionistic method of scoring.

REFERENCES

1. Buros O.K. (Ed) -The Fourth Mental Measurements Yearbook, New Brunswick, Rutgers University Press 1953.

2. Cast B.M.D. - The efficiency of different methods of marking English composition. Brit J. Edu Psychol. 1939-40.

3. Cronbach L.J. - Test reliability, its meaning and determination Psychometrika

4. Hartog P. - The marking of English Essays.

5. Hartog P & Rhodes E.C. - An Examination of Examinations.

6. Hartog P - The marks of Examiners - London Macmillan 193(

7. Kamat V.M.- The improvement of examination J. Educ. and Psychol 1954 203-212.

8. Taylor P.H. - The effects of different instruction to markers in the marking of English Essay. Brit J. Educational Psychol. 1959 29, 70-80.

9. Sims V.M. - Improving the measuring qualities of an essay examination J. Educ. Res, 1933. 27, 20-31.

10. Sims V.M. - Reducing the variability of Essay examination marks through eliminating variations i standards of grading J. Educ. Res, 1933 26, 637-647.

11. Edwards Penfold D.M. - Essay marking experiments. Shorter and longer essays. Brit J. Educ. Psychol 1956, 26, 123-126.

12. Dr.H.J.Taylor - The question paper.

13. Specimen test items for Secondary Schools (social Studies) (Directorate of Extension Programmes for Secondary Education).

14. Engelhart M.D.- Examination, Encyclopedia of edu. Research, (Ed: Monroe.W.S) New York Macmillan 1952.

15. Cureton E.E. - The difinition & estimation of test reliability Educ. Psychol Measurement 1958, 18, 715-738.

16. Vernon P.E. and G.D. Millican : Further Study of reliability of English essay. Brit J. Stat Psychology 1954, 7, 65-74.

17. Dutta A.D. - The reliability of Essay mark Ind J. Psychole. 1927., 2, 127-130.

18. Dutta A.D. - Objective examination. Ind. J. Psychol. 1928, 3, 613, 714.

19. Basumallick. Tapon - Reliability of Essay type Examination J of Edu & Psychol Oct. 1959, XVII 3, 127 - 156.

20. Lamb.H. - A comparison of two methods of marking (Summary of Thesis) Brit J. Educ. Psychology. 1953, 23, 131-33

21. Adkins D.C. - Construction and analysis of Achievement tests, Washington, Govt. Printing Office 1947.

22. Mathur R.K. - Effectiveness of board and university examination in India, Bulletin of the Institute of Public Administration, Patna University, Vol.7. No.1. April 1962.

23. Gyan A.K. & others: - Measurement of achievement in English, Report No.-II.

APPENDICES

Table 1.

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	27	25	19	24	23
2	18	24	15	17	14
3	19.0	24	15	17	14
4	15	20	14	12	7
5	21.0	21	10	19	10

Table 1.1. (Analysis of Variance Table)

Source of Variation	df	ss	ms	F
Between student	4	269.6	67.4	
Within student	20	374.4	18.72	$F = \frac{69.2}{6.1}$
Between Examiner	4	276.8	69.2	$= 11.34$
Residual	16	97.6	6.1	
Total	24	644		

Intraclass correlation = .35

Table 1.2 (Adjusted Data).

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	24.8	20	21.8	24.2	27.2
2	15.8	19	17.8	17.2	18.2
3	16.8	19	19.8	16.2	18.2
4	12.8	15	16.8	12.2	11.2
5.	18.8	16	12.8	19.2	14.2

Table 1.3. (Analysis of Variance Table)

Source of Variation	df	ss	ms
Between Student	4	269.6	67.4
Within student	16	97.4	6.09
Total:	20	367.0	

Intraclass correlation = .68

Table 1.4.

Students	Examiners E_1	E_2	E_3	E_4	E_5	Sum of ranks	Square of sum of ranks.
1	1	1	1	1	1	5	25
2	4	2.5	3	3	2.5	15	225
3	3	2.5	2	4	2.5	14	196
4	5	5	4	5	5	24	576
5	2	4	5	2	4	17	289
						75	1311

Co-efficient of Concordence = .744

Group 2

Table 2.

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	22	15	19	16	16
2	17	12	15	15	17
3	20	15	18	17	13
4	21	18	20	17	19
5	14	13	14	13	12

Table 2.1 (Analysis of Variance Table).

Sources of Variation	df	ss	ms	F
Between Student	4	99.44	24.86	
Within Student	20	92	4.6	$F = \frac{14.05}{2.24}$
Between Examiner	4	56.24	14.06	$= 6.276$
Residual	16	35.76	2.24	
Total:	24	191.44		

Intraclass correlation = .47

Table 2.2. (Adjusted Data)

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	19.52	16.72	18.12	16.72	16.92
2	14.52	13.72	14.12	15.72	17.92
3	17.52	16.72	17.12	17.72	13.92
4	18.52	19.72	19.12	17.72	19.92
5	11.52	14.72	13.12	13.72	12.92

Table 2.3 (Analysis of Variance Table)

Sources of Variation	df	ss	ms
Between Student	4	99.44	24.86
Within Student	16	35.75	2.234
Total:	20	135.19	

Intraclass Correlation = .669

Table 2.4.

Students	Examiners E_1	E_2	E_3	E_4	E_5	Sum of ranks	Square of sum of rnks
1	1	2.5	2	3	3	11.5	132.25
2	4	5	4	4	2	19.0	361.00
3	3	2.5	3	1.5	4	14.0	196.00
4	2	1	1	1.5	1	6.5	42.25
5	5	4	5	5	5	24.0	576
						75.0	1307.5

Co-efficient of Concordance = .728

Group 3.

Table 3.

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	16	20	20	5	6
2	12.5	26	19	12	19
3	16	29	22	12	20
4	22.5	30	17	11	19
5	19	29	19	13	16

Table 3.1. (Analysis of Variance Table)

Sources of Variation	df	ss	ms	F
Between Student	4	147.50	36.875	
Within Student	20	858.00	42.9	$F = \frac{173.5}{10.25}$
Between Examiner	4	694.00	173.5	= 16.8
Residual	16	164.00	10.25	
Total:	24	1005.50		

Intraclass correlation = -.03

Group 3 (Contd.)

Table 3.2. (Adjusted Data).

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	16.8	11.2	18.6	12.4	8
2	13.3	17.2	17.6	19.4	21
3	16.8	20.2	20.6	19.4	22
4	23.3	21.2	15.6	18.4	21
5	19.8	20.2	17.6	20.4	18

Table 3.3 (Analysis of Variance Table)

Sources of Variation	df	ss	ms
Between student	4	147.5	36.875
Within student	16	163.7	10.23
Total:	20	311.2	

Intraclass Correlation = .344

Table 3.4.

Rank given by Examiners

	E_1	E_2	E_3	E_4	E_5	Sum of ranks	Square of sum of ranks
1	3.5	5	2	5	5	20.5	420.25
2	5	4	3.5	2.5	2.5	17.5	306.25
3	3.5	2.5	1	2.5	1	10.5	110.25
4	1	1	5	4	2.5	13.5	182.25
5	2	2.5	3.5	1	4	13	169.00
						75.0	1178.00

Co-efficient of Concordance = .212

Group 4.

Table 4.

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	6	10	7	8	9
2	11	18.5	9	14	9.5
3	8.5	15	9	11	12
4	12	14	10	13	7.5
5	15	17	15	16	6

Table 4.1 (Analysis of Variance Table)

Sources of Variation	df	ss	ms	
Between student	4	91.94	22.985	
Within student	20	229.25	11.46	$F=\frac{28.435}{7.22}$
Between Examiner	4	113.74	28.435	= 3.924
Residual	16	115.51	7.22	
Total:	24	321.19		

Intraclass Correlation = .17

Table 4.2. (Adjusted Data)

Students	Examiners E_1	E_2	E_3	E_4	E_5
1	6.82	6.42	8.32	6.92	11.52
2	11.82	14.92	10.32	12.92	12.02
3	9.32	11.42	10.32	9.92	14.52
4	12.82	10.42	11.32	11.92	10.02
5	15.82	13.42	16.32	14.92	8.52

Table 4.3. (Analysis of Variance Table)

Sources of Variance	df	ss	ms
Between student	4	91.94	23.185
Within Student	16	70.76	4.42
Total:	20	162.7	

Intraclass correlation = .46

Table 4.4.

Students	Rank given by Examiners					Sum of ranks	Square of sum of ranks
	E_1	E_2	E_3	E_4	E_5		
1	5	5	5	5	3	23	529
2	3	1	3,5	2	2	11.5	132.25
3	4	3	3.5	4	1	15.5	240.25
4	2	4	2	3	4	15	225.00
5	1	2	1	1	5	10	100.00
						75	1226.50

Co-efficient of Concordance = .406

UNSTRUCTURED QUESTIONS

All the three questions should be answered. Each question carries 15 marks.

Time: 1½ hours

1. What influence did Muslim contact have on Hinduism?
2. Discuss the causes and nature of risings of 1857.
3. Evaluate critically the Chief reforms of cornwallis.

STRUCTURED QUESTIONS

All questions should be answered. Each question carries 15 marks.

Time: 1½ hours.

1. What influence did Muslim contact have on Hinduism. Discuss under the following headings

   (a) Art, Music & Paining
   (b) Architecture (Buildings)
   (c) Dress and customs (culture)
   (d) Literature
   (e) Religion (worship)

2. Discuss the causes and nature of the risings of 1857. Under the following headings

   (a) social
   (b) Religious
   (c) Financial
   (d) Military
   (e) Military revolt or Ist war of independence

3. Evaluate critically the chief reform of cornwallis under the following headings.

   (a) Reform of public services
   (b) judicial reforms
   (c) Police reforms
   (d) Permanent settlement of Bengal (Land revenue reform)

Directions Given to Examiners.

Please try to follow the following directions:-

1. Please do not put any mark what so ever on the answer script.
2. Think of a model answer before examining the script.
3. Read all answers to one question before going on to the next
4. Do not give stress on language. Do not deduct marks for spelling mistakes, punctuation etc.
5. Avoid giving negative marking.
6. Please keep in mind the different weightage given to different parts of the questions as given below.

Q.1. What influence did Muslim contact had on Hinduism.

Art music & painting	- 3 marks
Architecture (Buildings)	- 3 marks
Dress & customs (Culture)	- 3 marks
Literature	3 marks
Religion	3 marks

Q.2. Discuss the causes & nature of the rising of 1857.

Social -	3 marks
Religious	3 marks
Financial	2 marks
Military	2 marks

Military revolt om 1st war of independance - 5 marks

Q.3. Evaluate critically the chief reforms of Cornwall

Reform of public services -	3
Judicial reforms -	3
Police reforms -	2
Permanent settlement of Bengal-	7

7. Each question carries 15 marks.

National Institute of Education
BRARY & DOCUMENTATION
Unit (N.C.E.R.T.)
Acc. No. 2112